Title:

Rare variants of large effect in *BRCA2* and *CHEK2* affect risk of lung cancer

Authors:

Yufei Wang[1,61], James D McKay[2,61,62], Thorunn Rafnar[3], Zhaoming Wang[4], Maria N Timofeeva[2], Peter Broderick[1], Xuchen Zong[5], Marina Laplana[6], Yongyue Wei[7], Younghun Han[8], Amy Lloyd[1], Manon Delahaye-Sourdeix[2], Daniel Chubb[1], Valerie Gaborieau[2], William Wheeler[9], Nilanjan Chatterjee[4], Gudmar Thorleifsson[3], Patrick Sulem[3], Geoffrey Liu[10], Rudolf Kaaks[11,12], Marc Henrion[1], Ben Kinnersley[1], Maxime Vallée[2], Florence LeCalvez-Kelm[2], Victoria L Stevens[13], Susan M Gapstur[13], Wei V Chen[14], David Zaridze[15], Neonilia Szeszenia-Dabrowska[16], Jolanta Lissowska[17], Peter Rudnai[18], Eleonora Fabianova[19], Dana Mates[20], Vladimir Bencko[21], Lenka Foretova[22], Vladimir Janout[23], Hans E Krokan[24], Maiken Elvestad Gabrielsen[24], Frank Skorpen[25], Lars Vatten[26], Inger Njølstad[27], Chu Chen[28], Gary Goodman[28], Simone Benhamou[29], Tonu Vooder[30], Kristjan Välk[31], Mari Nelis[32,33], Andres Metspalu[32], Marcin Lener[34], Jan Lubiński[34], Mattias Johansson[2], Paolo Vineis[35,36], Antonio Agudo[37], Francoise Clavel-Chapelon[38–40], H Bas Bueno-de-Mesquita[35,41,42], Dimitrios Trichopoulos[43–45], Kay-Tee Khaw[46], Mikael Johansson[47], Elisabete Weiderpass[48–51], Anne Tjønneland[52], Elio Riboli[35], Mark Lathrop[53], Ghislaine Scelo[2], Demetrius Albanes[4], Neil E Caporaso[4], Yuanqing Ye[54], Jian Gu[54], Xifeng Wu[54], Margaret R Spitz[55], Hendrik Dienemann[12,56], Albert Rosenberger[57], Li Su[7], Athena Matakidou[58], Timothy Eisen[59,60], Kari Stefansson[3], Angela Risch[6,12], Stephen J Chanock[4], David C Christiani[7], Rayjean J Hung[5], Paul Brennan[2], Maria Teresa Landi[4,61,62], Richard S Houlston[1,61,62] & Christopher I Amos[8,61,62]

Author affiliations:
1 Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, Surrey, UK.
2 International Agency for Research on Cancer (IARC, World Health Organization (WHO)), Lyon, France.
3 deCODE Genetics, Amgen, Reykjavik, Iceland.
4 Department of Health and Human Services, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health (NIH), Bethesda, Maryland, USA.
5 Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Ontario, Canada.
6 Division of Epigenomics and Cancer Risk Factors, German Cancer Research Center (DKFZ), Heidelberg, Germany.
7 Department of Environmental Health, Harvard School of Public Health, Boston, Massachusetts, USA.
8 Center for Genomic Medicine, Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Lebanon, New Hampshire, USA.
9 Information Management Services, Inc., Rockville, Maryland, USA.
10 Princess Margaret Hospital, University Health Network, Toronto, Ontario, Canada.
11 Division of Cancer Epidemiology, DKFZ, Heidelberg, Germany.
12 Translational Lung Research Center Heidelberg (TLRC-H), Member of the German Center for Lung Research (DZL), Heidelberg, Germany.

13 Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, USA.
14 Department of Genetics, University of Texas MD Anderson Cancer Center, Houston, Texas, USA.
15 Institute of Carcinogenesis, Russian N.N. Blokhin Cancer Research Centre, Moscow, Russia.
16 Department of Epidemiology, Institute of Occupational Medicine, Lodz, Poland.
17 The M. Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Warsaw, Poland.
18 National Institute of Environmental Health, Budapest, Hungary.
19 Regional Authority of Public Health, Banská Bystrica, Slovak Republic.
20 National Institute of Public Health, Bucharest, Romania.
21 1st Faculty of Medicine, Institute of Hygiene and Epidemiology, Charles University in Prague, Prague, Czech Republic.
22 Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech Republic.
23 Palacky University, Olomouc, Czech Republic.
24 Department of Laboratory Medicine, Children's and Women's Health, Norwegian University of Science and Technology, Trondheim, Norway.
25 Department of Laboratory Medicine, Children's and Women's Health, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway.
26 Department of Public Health and General Practice, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway.
27 Department of Community Medicine, University of Tromsø, Tromsø, Norway.
28 Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.
29 INSERM U946, Paris, France.
30 Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia.
31 Department of Biomedicine, University of Bergen, Bergen, Norway.
32 Estonian Genome Center, Institute of Molecular and Cell Biology, Tartu, Estonia.
33 Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland.
34 Department of Genetics and Pathology, International Hereditary Cancer Center, Pomeranian Medical University, Szczecin, Poland.
35 Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London, UK.
36 Unit of Molecular and Genetic Epidemiology, HuGeF Foundation, Torino, Italy.
37 Unit of Nutrition, Environment and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology, Barcelona, Spain.
38 INSERM, Centre for Research in Epidemiology and Population Health (CESP), U1018, Nutrition, Hormones and Women's Health Team, Villejuif, France.
39 Université Paris Sud, UMRS 1018, Villejuif, France.
40 Institut Gustave Roussy, Villejuif, France.
41 National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands.
42 Department of Gastroenterology and Hepatology, University Medical Centre, Utrecht, The Netherlands.
43 Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA.
44 Bureau of Epidemiologic Research, Academy of Athens, Athens, Greece.
45 Hellenic Health Foundation, Athens, Greece.

46 University of Cambridge School of Clinical Medicine, Clinical Gerontology Unit, Addenbrooke's Hospital, Cambridge, UK.
47 Department of Radiation Sciences, Umeå Universitet, Umeå, Sverige, Sweden.
48 Department of Community Medicine, Faculty of Health Sciences, University of Tromsø, Tromsø, Norway.
49 Department of Research, Cancer Registry of Norway, Oslo, Norway.
50 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.
51 Samfundet Folkhälsan, Helsinki, Finland.
52 Danish Cancer Society Research Center, Copenhagen, Denmark.
53 Centre d'Etude du Polymorphisme Humain (CEPH), Paris, France.
54 Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA.
55 Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas, USA.
56 Department of Thoracic Surgery, Thoraxklinik at University Hospital Heidelberg, Heidelberg, Germany.
57 Department of Genetic Epidemiology, University of Göttingen, Göttingen, Germany.
58 Cancer Research UK Cambridge Institute, Li Ka Shing Centre, Cambridge, UK.
59 Department of Oncology, Cambridge University Hospitals National Health Service Foundation Trust, Cambridge, UK.
60 Addenbrooke's Hospital, Cambridge Biomedical Campus, Cambridge, UK.
61 These authors contributed equally to this work.
62 These authors jointly directed this work.

Corresponding author:
Richard S Houlston, Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, Surrey, UK. E-Mail: richard.houlston@icr.ac.uk
Maria Teresa Landi. Department of Health and Human Services, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health (NIH), Bethesda, Maryland, USA. E-Mail: landim@mail.nih.gov

Abstract

We conducted imputation to the 1000 Genomes Project of four genome-wide association studies of lung cancer in populations of European ancestry (11,348 cases and 15,861 controls) and genotyped an additional 10,246 cases and 38,295 controls for follow-up. We identified large-effect genome-wide associations for squamous lung cancer with the rare variants BRCA2 p.Lys3326X (rs11571833, odds ratio (OR) = 2.47, $P = 4.74 \times 10{-}20$) and CHEK2 p.Ile157Thr (rs17879961, OR = 0.38, $P = 1.27 \times 10{-}13$). We also showed an association between common variation at 3q28 (*TP63*, rs13314271, OR = 1.13, $P = 7.22 \times 10{-}10$) and lung adenocarcinoma that had been previously reported only in Asians. These findings provide further evidence for inherited genetic susceptibility to lung cancer and its biological basis. Additionally, our analysis demonstrates that imputation can identify rare disease-causing variants with substantive effects on cancer risk from preexisting genome-wide association study data.

Introduction

Lung cancer causes over 1 million deaths each year worldwide1. Although primarily caused by tobacco smoking, studies have also implicated inherited genetic factors in the etiology of lung cancer; notably, genome-wide association studies (GWAS) in Europeans have consistently identified polymorphic variation at 15q25.1 (*CHRNA5-CHRNA3-CHRNB4*), 5p15.33 (*TERT-CLPTM1*) and 6p21.33 (*BAG6* (also called *BAT3*)-*MSH5*) as determinants of lung cancer risk2–6. Additionally, susceptibility loci for lung cancer at 3q28, 6q22.2, 13q12.12, 10q25.2 and 22q12.2 in Asians have been identified through GWAS7–9.

Non–small cell lung cancer (NSCLC) is the most common lung cancer histology, comprised primarily of adenocarcinoma (AD) and squamous cell carcinoma (SQ). These lung cancer histologies have different molecular characteristics that reflect differences in etiology and carcinogenesis10. Perhaps not surprisingly, there is variability in the genetic effects on lung cancer risk by histology, with subtype-specific associations at 5p15.33 (*TERT-CLPTM1*) for AD11,12 and at 9p21 (*CDKN2A/CDKN2B*)13 and 12q13.33 (*RAD52*)14 for SQ. In addition, the 6p21.33 associations are stronger for SQ than for AD13.

To identify additional lung cancer susceptibility loci, we conducted a meta-analysis of four lung cancer GWAS in populations of European ancestry: the MD Anderson Cancer Center (MDACC) GWAS, the Institute of Cancer Research (ICR) GWAS, the National Cancer Institute (NCI) GWAS and the International Agency for Research on Cancer (IARC) GWAS (Online Methods), which were genotyped using Illumina HumanHap

317, 317+240S, 370Duo, 550, 610 or 1M arrays (**Supplementary Table 1**). After filtering, the studies provided genotypes on 11,348 cases and 15,861 controls (**Supplementary Table 1**). Before undertaking meta-analysis of the GWAS data, we searched for potential errors and biases in the data sets. Quantile-quantile (Q-Q) plots of genome-wide association test statistics showed minimal inflation, rendering substantial cryptic population substructure or differential genotype calling between cases and controls unlikely (λ = 1.01–1.05; **Supplementary Fig. 1**). To bring genotype data obtained from different arrays into a common platform and recover untyped genotypes, we imputed >10 million SNPs using 1000 Genomes Project data as the reference. Q-Q plots for all SNPs and those restricted to rare SNPs (minor allele frequency (MAF) <1%) after imputation did not show evidence of substantive overdispersion introduced by imputation (λ = 0.99–1.06 and λ = 0.82–1.05, respectively; **Supplementary Fig. 1**).

Pooling data from each GWAS, we derived joint ORs and 95% confidence intervals (CIs) under a fixed-effects model for each SNP and the associated per-allele *P* values. To explore variability in associations according to tumor histology, we derived ORs for all lung cancer, AD and SQ.

Our meta-analysis identified 50 SNPs that showed evidence of association with lung cancer, AD or SQ ($P < 5.0 \times 10{-}6$; **Fig. 1**) at loci not reported previously in Europeans (**Fig. 1**). We evaluated 1-Mb regions encompassing these 50 SNPs for association through *in silico* replication in the Harvard15 and deCODE16 series. Nine of the SNPs within these 50 regions showed support for an association (combined $P < 5.0 \times 10{-}7$). We attempted genotyping of these nine SNPs in four additional series: the Heidelberg–European Prospective Investigation into Cancer and Nutrition (EPIC), ICR, IARC and Toronto replications (**Supplementary Table 2b** and Online Methods). rs185577307

could not be genotyped because of repetitive sequence. Collectively, genotypes were available from 21,594 cases and 54,156 controls, providing 80% power to detect a variant with MAF of 0.01 and conferring a relative risk of ≥1.5. In the combined analysis of all GWAS plus replication series data, SNPs mapping to 13q13.1 (rs11571833 and rs56084662), 22q12.1 (rs17879961) and 3q28 (rs13314271) showed evidence for association, which was statistically significant after adjustment for multiple testing ($P < 3.0 \times 10{-}9$; **Fig. 2** and **Supplementary Table 3**). We confirmed the high fidelity of imputation by genotyping rs11571833, rs17879961 and rs13314271 in subsets of the ICR, IARC, NCI and MDACC GWAS (**Supplementary Table 2** and Online Methods). The NCI GWAS comprised samples from Finland, Italy and the United States. The IARC GWAS comprised samples from ten series from western and eastern Europe and the United States. Although adjustment of test statistics for principal components generated on common SNPs had been applied to these GWAS, confounding of rare variants in spatially structured populations is not necessarily corrected by such methods17. We therefore investigated whether country of origin had an impact on the associations at 13q13.1 and 22q12.1; the associations remained statistically highly significant ($P < 5.0 \times 10{-}8$; **Supplementary Table 4**).

rs11571833 and rs56084662, localizing to 13q13.1 near or within *BRCA2*, are rare (MAF < 0.01), map 103 kb apart (32,972,376 bp and 32,869,614 bp, respectively) and are moderately correlated ($r2 = 0.45$ and $D' = 0.82$ based on genotypes from the Heidelberg-EPIC, IARC, ICR and Toronto replication series; **Fig. 3**). rs11571833 (c.9976A>T) is responsible for BRCA2 p.Lys3326X, whereas rs56084662 is located in the 3′ UTR of *FRY*. Although the association provided by rs11571833 was substantially stronger than that provided by rs56084662 in the combined analysis (OR = 1.83, $P = 2.11$

× 10−19 and $P = 1.88 \times 10^{-15}$, respectively), a conditional analysis based on directly genotyped samples in the replication series was consistent with the two SNPs tagging the same haplotype. The association at rs11571833 is driven primarily by a relationship with SQ histology rather than AD histology (OR = 2.47, $P = 4.74 \times 10^{-20}$ and OR = 1.47, $P = 4.66 \times 10^{-4}$, respectively; **Fig. 2** and **Supplementary Table 3**). A stronger role for *BRCA2* in SQ etiology than in AD etiology is reflected in the higher observed mutational frequency in the respective lung cancers (~6% and 1% (refs. 18,19)). Thr9976 was recently shown to confer a 1.26-fold increased breast cancer risk20 and has been suggested previously as a risk factor for esophageal and pancreatic cancers21,22. We found no evidence for an association between Thr9976 and lung cancer risk in nonsmokers using directly genotyped samples (**Supplementary Table 2**); however, these cases comprised <10% of each cohort, and therefore our power to demonstrate a relationship was limited. Previous analyses of families carrying highly penetrant *BRCA2* mutations have found either no evidence for any excess risk or a reduced risk of lung cancer in carriers23,24. A possible explanation for these observations is that members of the families studied tended to smoke less than the general population24.

The RAD51-BRCA2 interaction is pivotal for BRCA2-mediated double strand–break repair, and exon 27 of *BRCA2* encodes one of the highly conserved RAD51 binding domains: homozygous deletion of exon 27 in mice confers susceptibility to tumors, including lung cancer25. Thr9976 leads to the loss of the C-terminal domain of BRCA2, inviting speculation that the SNP is functional. Although the deleted region is distal to the RAD51 binding domain and an impact on nuclear localization is unknown26,27, the nearby BRCA2 p.Thr3387Ala alteration interrupts CHK2 phosphorylation and abrogates BRCA2-CHK2-RAD51–mediated recombination repair28. Alternatively, the association

might be a consequence of linkage disequilibrium (LD) with another *BRCA2* mutation. Studies of families with breast cancer of northern European ancestry show that the *BRCA2* c.6275delTT and c.4889C>G mutations, which are highly penetrant for breast and ovarian cancer, originated on a p.Lys3326X haplotype29. To gain further insight into a probable genetic basis of the 13q13.1 lung cancer association, we sequenced germline DNA from 70 individuals with lung cancer who carried c.9976A>T from the UK Genetic Lung Cancer Predisposition Study for the c.6275delTT and c.4889C>G mutations; we did not find c.6275delTT or c.4889C>G in any of these individuals. Similarly, sequencing the coding region of *BRCA2* identified no clearly pathogenic mutations among 13 individuals from the 1958 British Birth Cohort (58BC), 11 individuals with lung cancer from IARC or 24 individuals with lung cancer carrying Thr9976 from TCGA. In Iceland, Thr9976 is not correlated with the founder *BRCA2* mutation resulting in p.256_257del (c.999del5), which greatly increases the risk of breast and ovarian cancer. Paradoxically, whereas Thr9976 is a risk factor for lung cancer, in this population this SNP is not associated with risk of breast or ovarian cancer (**Supplementary Table 5**). Although *in vitro* studies have failed to demonstrate that p.Lys3326X affects DNA repair30, our findings raise the possibility that p.Lys3326X may have a direct effect on lung cancer risk. The fact that somatic mutation of *BRCA2* is not associated with p.Lys3326X carrier status19 (**Supplementary Table 6a**) suggests that any impact the SNP has on lung cancer risk is mediated through alternative mechanisms.

The relationship at 22q12.1 between the rs17879961 (c.470T>C) and SQ in the combined series (OR = 0.38, $P = 1.27 \times 10{-}13$) validates an association that has been reported previously31,32 (**Fig. 2** and **Supplementary Tables 3** and **4**). The frequency of rs17879961 varies markedly between populations: it has a MAF of ~5% in eastern Europeans (for example, individuals in the IARC series) but is almost monomorphic in

most northern Europeans. This likely accounts for the failure to demonstrate a significant relationship in the ICR, MDACC, Toronto and deCODE series, which comprise largely western European populations (**Fig. 2** and **Supplementary Table 3**). rs17879961 is responsible for the missense mutation in *CHEK2* resulting in p.Ile157Thr; *CHEK2* is a cell cycle–control gene encoding a pluripotent kinase that can cause arrest or apoptosis in response to DNA damage. Acquired mutation of *CHEK2* is rarely seen in lung cancer, and the CHEK2 p.Ile157Thr alteration does not appear to correlate with mutation (**Supplementary Table 6a**), raising the possibility that carrier status *per se* influences cancer risk. The p.Ile157Thr substitution lies in a functionally important domain of CHEK2 and causes reduced or abolished binding of principal substrates. Although Cys470 increases breast cancer risk33, here Cys470 was associated with reduced lung cancer risk. A mechanism for the paradoxical associations is not immediately apparent. However, *CHEK2* can have opposite effects on damaged stem cells, retarding stem cell division until DNA damage is repaired or activating apoptosis if damage cannot be repaired. Although speculative, in the presence of continued DNA damage to squamous epithelia by tobacco smoke, the normal stem cell defenses involving CHEK2 might be attenuated by a reduction in CHEK2 activity as a result of p.Ile151Thr31. Concordant with such a model is our observation of a paradoxically increased lung cancer risk in nonsmokers (P = 0.05) and in correlated subgroups of AD and women, although this increase was based on small numbers (**Supplementary Table 2**).

The association between variation at 3q28 marked by rs13314271 and lung cancer risk was restricted to AD (OR = 1.13, $P = 7.22 \times 10{-}10$; **Fig. 2** and **Supplementary Table 3**). rs13314271 maps within intron 1 of *TP63* (**Fig. 3**). Variation at *TP63* defined by the intron 1 SNP rs4488809, which is in complete LD with rs13314271 ($r2 = 1.00$, $D' = 1.00$),

is associated with AD in Asians8. Our findings provide robust evidence for the generalizability of a relationship between 3q28 variation and AD. We found a weak association between rs13314271 and lung cancer risk in nonsmokers (P = 0.03; **Supplementary Table 2b**). *TP63* is a member of the tumor suppressor *TP53* gene family, which is pivotal in cellular differentiation and responsiveness to cellular stress34,35. Exposure of cells to DNA damage leads to induction of *TP63*, and both isoforms have the ability to transactivate *TP53* target genes, thereby affecting cellular responsiveness to DNA damage36. Although rs13314271 does not map to an evolutionary conserved region, rs7636839, which is correlated with rs13314271 and rs4488809 ($r2$ = 1.0), does map to an evolutionarily conserved region and has predicted enhancer activity (**Supplementary Table 6b**). Moreover, rs4488809 has been shown to be an expression quantitative trait locus for *TP63* in lung tissue37. Although the mechanism by which 3q28 variation affects AD development is unknown, accumulation of DNA damage and a lack of response to genotoxic stress are recognized to contribute to lung carcinogenesis; hence, loss of repair fidelity as a consequence of differential *TP63* expression is likely deleterious.

There was no association between rs11571833, rs17879961 and rs13314271 genotypes and cigarette consumption on the basis of smoking information on 43,693 Icelandic subjects (**Supplementary Table 7**), which is in contrast to the association of 15q25 and risk of lung cancer.

Although there is some overlap, distinct DNA lesions are ostensibly repaired by different DNA repair pathways. Histology-specific relationships seen implicate the BRCA2-CHEK2-RAD52 double strand–break repair and homologous recombination pathways as

a determinant of SQ and defective TP53 and TERT apoptosis-telomerase regulation as a basis of AD risk. In conclusion, our findings provide further evidence for inherited genetic susceptibility to lung cancer and underscore the importance of searching for histology-specific risk variants. Our data also provide an important proof of principle that 1000 Genomes imputation can be used to detect new, low-frequency, large-effect associations, thereby extending the utility of preexisting GWAS data. Notably, this study facilitated the identification of BRCA2 Thr9976, which is the strongest genetic association in lung cancer reported so far. For a smoker carrying this variant (2% of the population), the risk of developing lung cancer is approximately doubled, which may have implications for identifying high-risk ever-smoking subjects for lung cancer screening. Additionally, future study of the effects of PARP inhibition in smokers with lung cancer carrying BRCA2 Thr9976 may be warranted.

URLs.

R suite, http://www.r-project.org/

1000 Genomes Project, http://www.1000genomes.org/

SNAP, http://www.broadinstitute.org/mpg/snap/

IMPUTE2, http://mathgen.stats.ox.ac.uk/impute/impute_v2.html

MACH, http://www.sph.umich.edu/csg/abecasis/MACH/

Minimac, http://genome.sph.umich.edu/wiki/Minimac/

SNPTEST, https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html

ProbABEL, http://www.genabel.org/packages/ProbABEL; mach2dat, http://genome.sph.umich.edu/wiki/Mach2dat:_Association_with_MACH_output

Wellcome Trust Case Control Consortium, http://www.wtccc.org.uk/

RegulomeDB, http://regulome.stanford.edu/

HaploReg v2, http://www.broadinstitute.org/mammals/haploreg/haploreg.php

Transdisciplinary Research In Cancer of the Lung (TRICL), http://u19tricl.org/

Genetic Associations and MEchanisms in ONcology (GAME-ON) consortium, http://epi.grants.cancer.gov/gameon/

International Lung Cancer Consortium (ILCCO), http://ilcco.iarc.fr/

Icelandic Cancer Registry, http://www.krabbameinsskra.is/

Genome Analysis Toolkit (GATK), http://www.broadinstitute.org/gatk/

The Cancer Genome Atlas (TCGA), http://cancergenome.nih.gov/

Leiden Open Variation Database (LOVD), http://www.lovd.nl/3.0/home/

Breast Cancer IARC database, http://brca.iarc.fr/.

Acknowledgments

We thank all individuals who participated in this study. We are also grateful to the patients, clinicians and allied health care professions. We thank Z. Chen and K. Boyle for sample handling and data management of the Toronto study, and L. Admas and L.R. Zhang for field recruitment. We thank L. Su, Y. Zhao, G. Liu, J. Wain, R. Heist and K. Asomaning for providing computing support at MDACC. We thank G. Thomas and Synergy Lyon Cancer (Lyon France) for high performance computing support and J.

Olivier and A. Chabrier for IARC's PGM ion torrent sequencing optimization and TaqMan genotyping, respectively. We thank D. Goldgar for sharing information from The Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA) on sequence variation in BRCA2 from familial breast cancer analysis. We acknowledge the Icelandic Cancer Registry (http://www.krabbameinsskra.is/indexen.jsp?id=summary) for assistance in the ascertainment of the Icelandic patients with lung cancer. The ICR study made use of genotyping data from the Wellcome Trust Case-Control Consortium 2 (WTCCC2); a full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. We acknowledge The Cancer Genome Atlas (TCGA) for their contribution of lung cancer genomic data to this study (TCGA Project Number 3230). We also acknowledge support from the National Institute for Health Research Biomedical Research Centre at the Royal Marsden Hospital. This study was supported by the NIH (U19CA148127, R01CA055769, 5R01CA127219, 5R01CA133996 and 5R01CA121197). The work performed at ICR was supported by Cancer Research UK (C1298/A8780 and C1298/A8362), National Cancer Research Network (NCRN), HEAL, Sanofi-Aventis and National Health Service funding to the Royal Marsden Hospital and Institute of Cancer Research, as well as the National Institute for Health Research Biomedical Research Centre. B.K. was the recipient of a Sir John Fisher Foundation PhD studentship. Work at ICR was also supported by NIH GM103534 and the Institute for Quantitative Biomedical Sciences at Dartmouth to C.I.A. The work performed in Toronto was supported by The Canadian Cancer Society Research Institute (020214), Ontario Institute of Cancer and Cancer Care Ontario Chair Award to R.J.H. and G.L. and the Alan Brown Chair and Lusi Wong Programs at the Princess Margaret Hospital Foundation. The work performed at Heidelberg was supported by Deutsche Krebshilfe (70-2387 and 70-2919) and the German Federal Ministry of Education and

Research (EPIC-Heidelberg). The work performed at IARC was supported by the Institut National du Cancer, France, the European Community (LSHG-CT-2005-512113), the Norwegian Cancer Association, the Functional Genomics Programme of Research Council of Norway, the European Regional Development Fund and the State Budget of the Czech Republic (RECAMO, CZ.1.05/2.1.00/03.0101), the NIH (R01-CA111703 and UO1-CA63673), the Fred Hutchinson Cancer Research Center, the US NCI (R01 CA092039), an FP7 grant (REGPOT 245536), the Estonian Government (SF0180142s08), the EU European Regional Development Fund in the frame of Centre of Excellence in Genomics and Estonian Research Infrastructure's Roadmap and the University of Tartu (SP1GVARENG) and an IARC Postdoctoral Fellowship (M.N.T.). Work at the NCI was supported by the Intramural Research Program of the NIH, the NCI, US Public Health Service contracts NCI (N01-CN-45165, N01-RC-45035, N01-RC-37004, NO1-CN-25514, NO1-CN-25515, NO1-CN-25512, NO1-CN-25513, NO1-CN-25516, NO1-CN-25511, NO1-CN-25524, NO1-CN-25518, NO1-CN-75022, NO1-CN-25476 and NO1-CN-25404), the American Cancer Society, the NIH Genes, Environment and Health Initiative in part by HG-06-033-NCI-01 and RO1HL091172-01, genotyping at the Johns Hopkins University Center for Inherited Disease Research (U01HG004438 and NIH HHSN268200782096C) and study coordination at the GENEVA Coordination Center (U01 HG004446). Work was also supported by NIH grants (P50 CA70907, R01CA121197, RO1 CA127219, U19 CA148127 and RO1 CA55769) and a Cancer Prevention Research Institute of Texas grant (RP100443). Genotyping was provided by the Center for Inherited Disease Research (CIDR). Work performed at Harvard was supported by the NIH (CA074386, CA092824 and CA090578). The Icelandic study was supported in part by NIH DA17932.

Author contributions

R.S.H. and Y. Wang conceived the study and provided overall project management and drafted the paper. In the UK, Y. Wang performed statistics and bioinformatics of UK data and conducted all meta-analyses; additional support was provided by M.H.; P. Broderick oversaw genotyping and sequencing; A.L. and B.K. performed genotyping and Sanger sequencing; A. Matakidou, T.E. and R.S.H. were responsible for the development and operation of the Genetic Lung Cancer Predisposition Study (GELCAPS); and D.C. and P. Broderick performed next-generation sequencing. At IARC, J.D.M. and P. Brennan provided overall project management; M.N.T., M.D.-S., V.G. and M.V. performed statistics and bioinformatics of IARC data and conducted meta-analysis; J.D.M. and F.L.-K. oversaw genotyping and sequencing; and G.S., D.Z., N.S.-D., J. Lissowska, P.R., E.F., D.M., V.B., L.F., V.J., H.E.K., M.E.G., F.S., L.V., I.N., C.C., G.G., M. Lathrop, S.B., T.V., K.V., M.N., A. Metspalu, M. Lathrop, J. Lubiński, Mattias Johansson, P.V., A.A., F.C.-C., H.B.-d.-M., D.T., K.-T.K., Mikael Johansson, E.W., A.T., R.K. and E.R. provided samples and data. For the Dartmouth and MDACC component, C.I.A. provided overall project management, obtained support for genotyping and contributed to statistical analyses; W.V.C. performed imputation analysis; Y.H. performed statistical analyses; and M.R.S. oversaw sample collection and development of the epidemiological studies. M.R.S. was also responsible for collecting samples that are a part of this research. X.W. provided ongoing support for the research protocol and supported large laboratory management of samples. Y.Y. and J.G. performed genotyping. At the NCI, M.T.L. was responsible for the overall project and managed the Environment and Genetics in Lung Cancer Etiology (EAGLE) study; N.E.C. managed the Prostate, Lung, Colon, Ovary Screening Trial (PLCO) study; D.A. managed the α-Tocopherol, β-Carotene Cancer Prevention Study (ATBC); S.M.G. and V.L.S. managed the Cancer Prevention Study II Nutrition Cohort (CPS-II) study; N.C. and W.W. performed statistical analyses; Z.W.

performed genotyping and imputation analysis; and S.J.C. oversaw genotyping and imputation analysis. At decode, T.R. and K.S. were responsible for the development and operation of deCODE's lung cancer study; and G.T. and P.S. performed the imputations and statistical analysis of the Icelandic data. At Harvard, D.C.C. was responsible for the overall conduct of the project; L.S. was responsible for sample management, genotyping and laboratory quality control; and Y. Wei performed data management and statistical analyses. For the Heidelberg-EPIC replication, M. Laplana managed DNA samples and performed genotyping; A. Rosenberger managed genotype and phenotype information; A. Risch supervised genotyping and data analysis; and R.K., A. Risch and H.D. conceived and managed studies that contributed samples. For the Toronto replication, R.J.H. and G.L. provided overall supervision of the study conduct, including study design, field recruitment, genotyping and statistical analysis; and X.Z. performed the statistical analysis.

Competing financial interests

The authors declare no competing financial interests.

Figures

Figure 1 Genome-wide P values plotted against their respective chromosomal positions. (a–c) All lung cancer (a), AD (b) and SQ (c). Shown are the genome-wide P values (two sided) obtained using the Cochran-Armitage trend test from analysis of 8.9 million successfully imputed autosomal SNPs in 11,348 cases and 15,861 controls from the discovery phase. The red and blue horizontal lines represent the significance thresholds of P = 5.0 × 10−8 and P = 5.0 × 10−6, respectively. Any region that contained at least one association signal better than P = 5.0 × 10−6 was selected for the in silico replication.

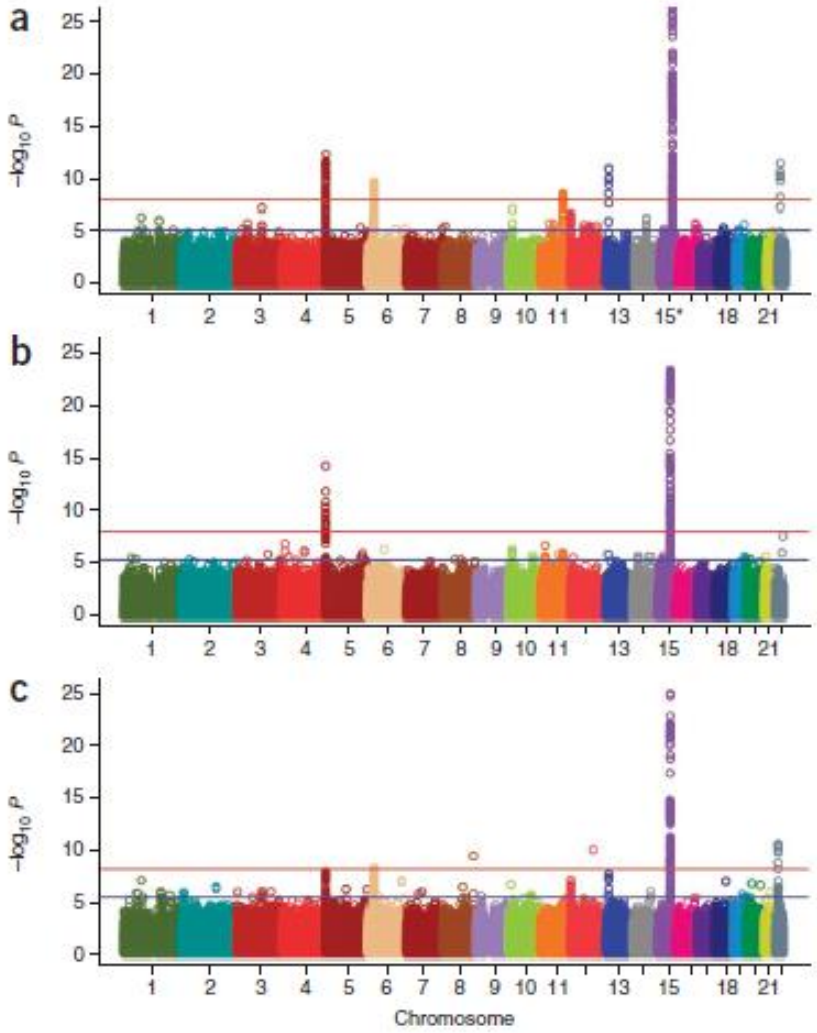


Figure 2 Plots of the ORs of lung cancer associated with 13q13.1 (rs11571833 and rs56084662), 22q12.1 (rs17879961) and 3q28 (rs13314271) risk loci. (a–l) All lung cancer based on 21,594 lung cancer cases and 54,156 controls (a–d), SQ based on 6,477 SQ cases and 53,333 controls (e–h) and AD based on 7,031 AD cases and 53,189 controls (i–l). The studies are weighted according to the inverse of the variance of the log of the OR calculated by unconditional logistic regression. Horizontal lines indicate the 95% CIs. Boxes are the OR point estimates, and the area of the box is proportional to the weight of the study. Diamonds and broken lines indicate the overall summary estimate derived under a fixed-effects (FE) model, with the CI given by the width. Unbroken vertical lines show the null value (OR = 1.0).

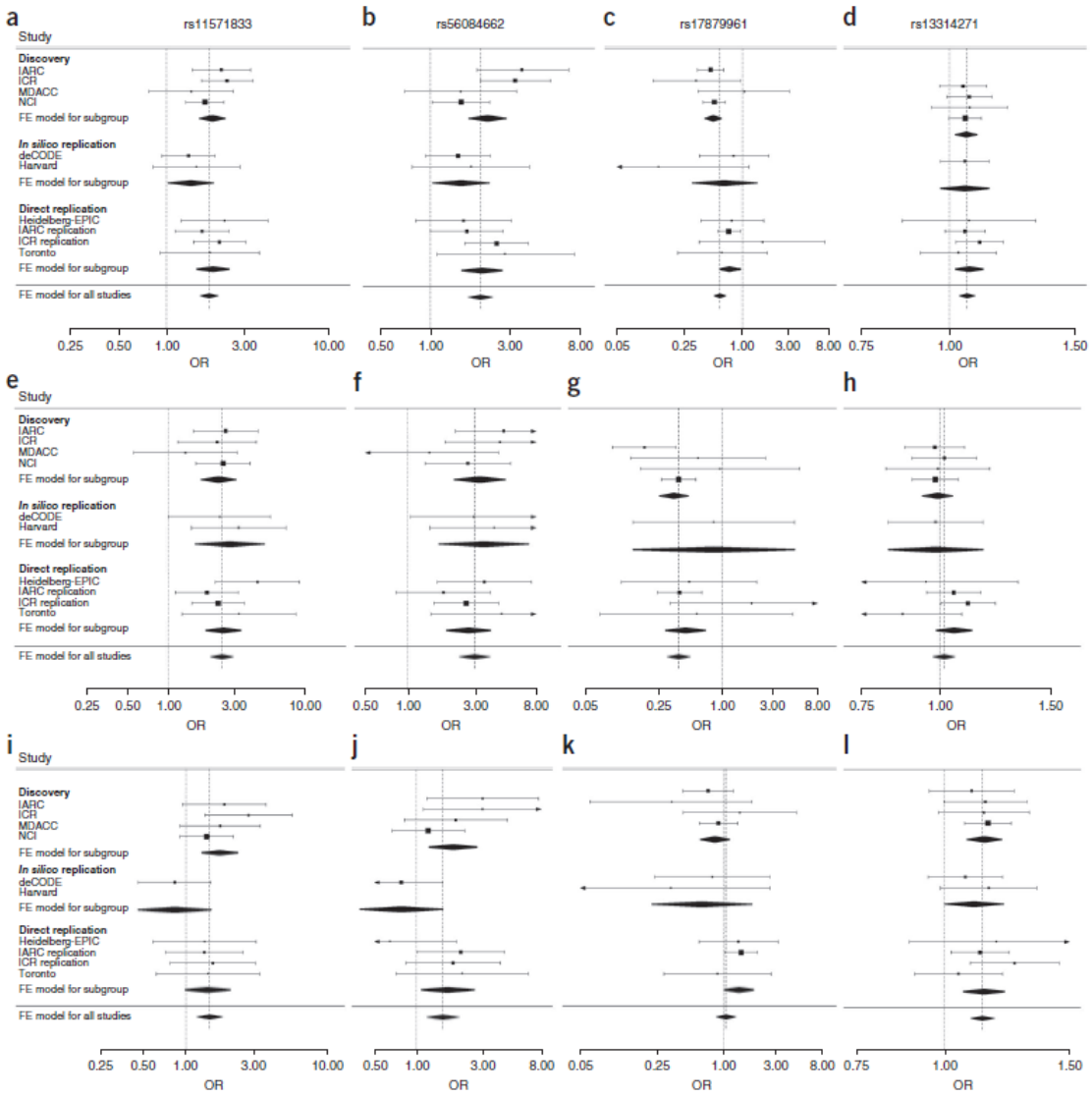
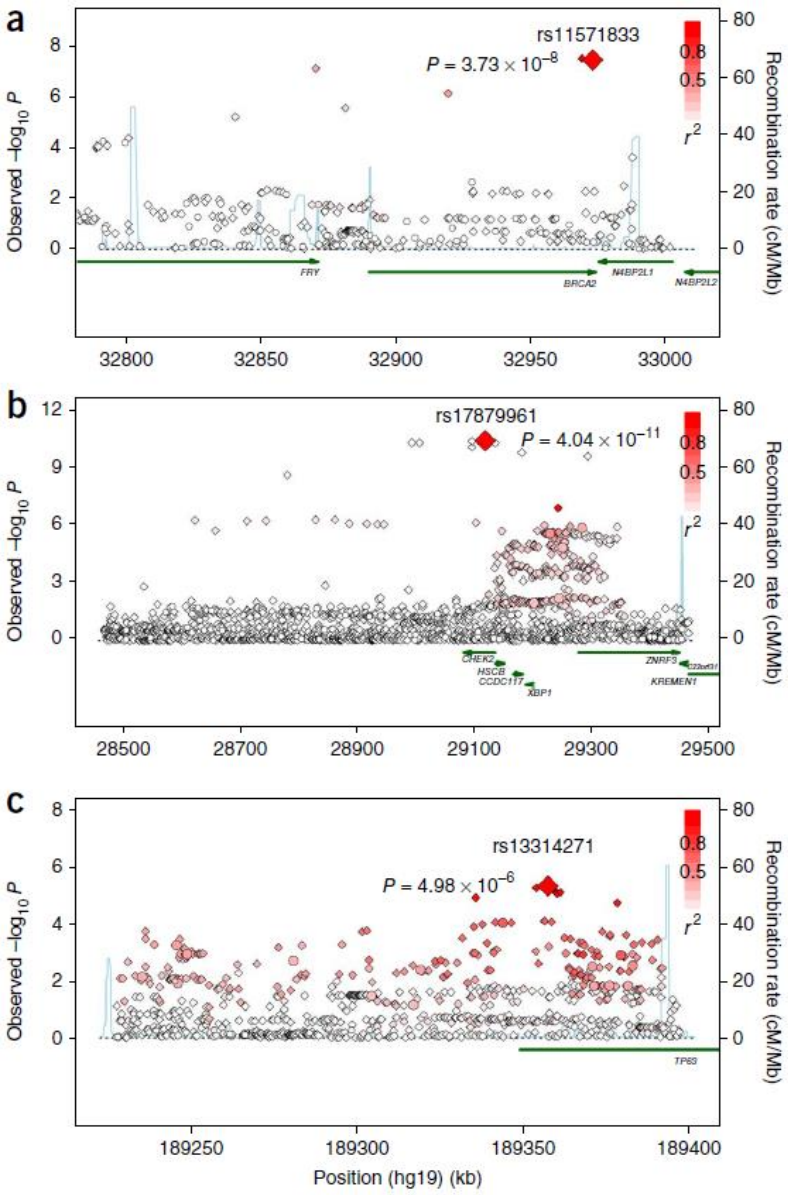


Figure 3 Regional plots of associations at susceptibility loci for SQ and AD. (a–c) Association results and recombination rates for the 13q13.1 in SQ (a), 22q12.1 in SQ (b) and 3q28 in AD (c). The SQ-related plots (a,b) were based on 3,275 SQ cases and 15,038 controls from the discovery phase; the AD-related plot (c) was based on 3,442 AD cases and 14,894 controls from the discovery phase. Association results of both genotyped (circles) and imputed (diamonds) SNPs in the GWAS samples and recombination rates for each locus are shown. For each plot, −log10 P values (y axes) of the SNPs are shown according to their chromosomal positions (x axes). The top genotyped SNP in each combined analysis is indicated by a large diamond and is labeled by its rsID. The color intensity of each symbol reflects the extent of LD with the top genotyped SNP: white (r2 = 0) through to dark red (r2 = 1.0). Genetic recombination rates (cM/Mb), estimated using HapMap CEU samples, are shown with a light blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative positions of genes and transcripts mapping to each region of association. Genes have been redrawn to show the relative positions; therefore, maps are not to physical scale.



References

1. Ferlay, J. et al. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int. J. Cancer 127, 2893–2917 (2010).

2. Hung, R.J. et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 452, 633–637 (2008).

3. Amos, C.I. et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat. Genet. 40, 616–622 (2008).

4. Thorgeirsson, T.E. et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 452, 638–642 (2008).

5. McKay, J.D. et al. Lung cancer susceptibility locus at 5p15.33. Nat. Genet. 40, 1404–1406 (2008).

6. Wang, Y. et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. Nat. Genet. 40, 1407–1409 (2008).

7. Hu, Z. et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. Nat. Genet. 43, 792–796 (2011).

8. Miki, D. et al. Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations. Nat. Genet. 42, 893–896 (2010).

9. Lan, Q. et al. Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. Nat. Genet. 44, 1330–1335 (2012).

10. Travis, W.D. et al. International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society: international multidisciplinary

classification of lung adenocarcinoma: executive summary. Proc. Am. Thorac. Soc. 8, 381–385 (2011).

11. Broderick, P. et al. Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. Cancer Res. 69, 6633–6641 (2009).

12. Landi, M.T. et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am. J. Hum. Genet. 85, 679–691 (2009).

13. Timofeeva, M.N. et al. Influence of common genetic variation on lung cancer risk: meta-analysis of 14,900 cases and 29,485 controls. Hum. Mol. Genet. 21,

4980–4995 (2012).

14. Shi, J. et al. Inherited variation at chromosome 12p13.33, including RAD52, influences the risk of squamous cell lung carcinoma. Cancer Discov. 2, 131–139 (2012).

15. Huang, Y.T. et al. Cigarette smoking increases copy number alterations in nonsmall-cell lung cancer. Proc. Natl. Acad. Sci. USA 108, 16345–16350 (2011).

16. Rafnar, T. et al. Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. Nat. Genet. 41, 221–227 (2009).

17. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. 44, 243–246 (2012).

18. Imielinski, M. et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell 150, 1107–1120 (2012).

19. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature 489, 519–525 (2012).

20. Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat. Genet. 45, 353–361 (2013).

21. Akbari, M.R. et al. Germline BRCA2 mutations and the risk of esophageal squamous cell carcinoma. Oncogene 27, 1290–1296 (2008).

22. Martin, S.T. et al. Increased prevalence of the BRCA2 polymorphic stop

codon K3326X among individuals with familial pancreatic cancer. Oncogene 24, 3652–3656 (2005).

23. Breast Cancer Linkage Consortium. Cancer risks in BRCA2 mutation carriers.

J. Natl. Cancer Inst. 91, 1310–1316 (1999).

24. van Asperen, C.J. et al. Cancer risks in BRCA2 families: estimates for sites other than breast and ovary. J. Med. Genet. 42, 711–719 (2005).

25. McAllister, K.A. et al. Cancer susceptibility of mice with a homozygous deletion in the COOH-terminal domain of the Brca2 gene. Cancer Res. 62, 990–994 (2002).

26. Spain, B.H., Larson, C.J., Shihabuddin, L.S., Gage, F.H. & Verma, I.M. Truncated BRCA2 is cytoplasmic: implications for cancer-linked mutations. Proc. Natl. Acad. Sci. USA 96, 13920–13925 (1999).

27. Yano, K. et al. Nuclear localization signals of the BRCA2 protein. Biochem. Biophys. Res. Commun. 270, 171–175 (2000).

28. Bahassi, E.M. et al. The checkpoint kinases Chk1 and Chk2 regulate the functional associations between hBRCA2 and Rad51 in response to DNA damage. Oncogene 27, 3977–3985 (2008).

29. Mazoyer, S. et al. A polymorphic stop codon in BRCA2. Nat. Genet. 14, 253–254 (1996).

30. Wu, K. et al. Functional evaluation and cancer risk assessment of BRCA2 unclassified variants. Cancer Res. 65, 417–426 (2005).

31. Brennan, P. et al. Uncommon CHEK2 mis-sense variant and reduced risk of tobacco-related cancers: case control study. Hum. Mol. Genet. 16, 1794–1801 (2007).

32. Cybulski, C. et al. Constitutional CHEK2 mutations are associated with a decreased risk of lung and laryngeal cancers. Carcinogenesis 29, 762–765 (2008).

33. Han, F.F., Guo, C.L. & Liu, L.H. The effect of CHEK2 variant I157T on cancer susceptibility: evidence from a meta-analysis. DNA Cell Biol. 32, 329–335 (2013).

34. Flores, E.R. The roles of p63 in cancer. Cell Cycle 6, 300–304 (2007).

35. Katoh, I., Aisaki, K.I., Kurata, S.I., Ikawa, S. & Ikawa, Y. p51A (TAp63γ), a p53 homolog, accumulates in response to DNA damage for cell regulation. Oncogene 19, 3126–3130 (2000).

36. Petitjean, A. et al. Properties of the six isoforms of p63: p53-like regulation in response to genotoxic stress and cross talk with ΔNp73. Carcinogenesis 29,

273–281 (2008).

37. Hao, K. et al. Lung eQTLs to help reveal the molecular underpinnings of asthma. PLoS Genet. 8, e1003029 (2012).

Online Materials & Methods

Studies.

The study was conducted under the auspices of the Transdisciplinary Research In Cancer of the Lung (TRICL) Research Team, which is a part of the Genetic Associations and MEchanisms in ONcology (GAME-ON) consortium and is associated with the International Lung Cancer Consortium (ILCCO). Tumors from patients were classified as AD, SQ, large-cell carcinoma (LCC), mixed adenosquamous carcinoma (MADSQ) and other NSCLC histologies following either the International Classification of Diseases for Oncology (ICD-O) or WHO coding. Tumors with overlapping histologies were classified as mixed.

Ethics.

All participants provided informed written consent. All studies were reviewed and approved by institutional ethics review committees at the involved institutions.

GWAS.

The meta-analysis was based on data from four previously reported lung cancer GWAS of European populations: the MDACC GWAS3, the ICR GWAS6, the NCI GWAS13 and the IARC GWAS2. In each of the studies, SNP genotyping had been performed using Illumina HumanHap 317, 317+240S, 370, 550, 610 or 1M arrays (Supplementary Table 1).

IARC GWAS. The IARC GWAS2 comprised 3,062 lung cancer cases and 4,455 controls derived from five case-control studies: (i) the Carotene and Retinol Efficacy Trial (CARET) cohort38; (ii) the Central Europe multicenter hospital-based case-control study39,40; (iii) the hospital-based case-control study from France40; (iv) the hospital-

based case-control lung cancer study from Estonia41,42; and (v) the population-based HUNT2/Tromsø IV lung cancer studies43. Patient and control DNAs were derived from EDTA–venous blood samples. The patients with lung cancer were classified according to ICD-O-3: SQ: 8070/3, 8071/3, 8072/3, 8074/3; AD: 8140/3, 8250/3, 8260/3, 8310/3, 8480/3, 8560/3, 8251/3, 8490/3, 8570/3, 8574/3; with tumors with overlapping histologies being classified as mixed. After applying standardized quality-control procedures, 2,533 cases and 3,791 controls were included in the current analysis (Supplementary Table 1).

NCI GWAS. Details of the NCI GWAS have been reported previously. Briefly, the study comprised samples from four series: (i) the Environment and Genetics in Lung cancer Etiology (EAGLE) study, a population-based case-control study of 2,100 lung cancer cases and 2,120 healthy controls enrolled in Italy between 2002 and 2005 (ref. 44), in which cancers were classified according to the ICD-O coding for histology and grading and histology of ~10% of tumors was confirmed by an independent pathologist from the NCI; (ii) the α-Tocopherol, β-Carotene Cancer Prevention Study (ATBC), a randomized primary prevention trial of 29,133 male smokers enrolled in Finland between 1985 and 1993 (ref. 45), in which ICD-O-2 and ICD-O-3 were used to classify tumors and cases diagnosed between 1985 and 1999 had histology reviewed by at least one pathologist (after 1999, histological coding (ICD-O-2 and ICD-O-3) was derived from the Finnish Cancer Registry); (iii) the Prostate, Lung, Colon, Ovary Screening Trial (PLCO), a randomized trial of 150,000 individuals enrolled in 10 US study centers between 1992 and 2001 (ref. 46), in which ICD-O-2 was used to classify tumors and quality assurance measures included reabstraction of 50 lung cancer diagnoses per year; and (iv) the Cancer Prevention Study II Nutrition Cohort (CPS-II), a cohort study of approximately 184,000 individuals enrolled by the American Cancer Society between 1992 and 1993 in 21 US

states, of which 109,379 provided a blood (36%) or buccal (64%) sample between 1998 and 2003 (refs. 12,47) and tumor histology was abstracted from Certified Tumor Registrars and coded using WHO ICD-O-2 and ICD-O-3. In this study, quality assurance was done by reabstracting 10% of all cancer diagnoses per year. After initial data quality control, the NCI GWAS included 5,739 cases and 5,848 controls; however, an additional 26 cases and 112 controls were excluded because of changes in case status and further quality-control filtering. The current meta-analysis included 5,713 lung cancer cases and 5,736 controls from the NCI GWAS (Supplementary Table 1).

ICR GWAS. The ICR GWAS comprised 1,952 cases (1,166 male; mean age at diagnosis 57 years, s.d. 6 years) with pathologically confirmed lung cancer ascertained through the Genetic Lung Cancer Predisposition Study (GELCAPS) conducted between March 1999 and July 2004 (ref. 48). All cases were British residents and were self-reported to be of European ancestry. To ensure that data and samples were collected from bona fide lung cancer cases and avoid issues of bias from survivorship, only incident cases with histologically or cytologically (if not AD) confirmed primary disease were ascertained. Tumors from patients were classified according to ICD-O3: specifically, SQ: 8070/3, 8071/3, 8072/3, 8074/3; AD: 8140/3, 8250/3, 8260/3, 8310/3, 8480/3, 8560/3, 8251/3, 8490/3, 8570/3, 8574/3; with tumors with overlapping histologies being classified as mixed. Patient DNA was derived from EDTA–venous blood samples using conventional methodologies. Genotype frequencies were compared with publicly accessible data generated by the UK Wellcome Trust Case-Control Consortium 2 (WTCCC2) study49 of individuals from the 1958 British Birth Cohort (58BC), and blood service was typed using Illumina Human1.2M-Duo Custom_v1 Array BeadChips.

MDACC GWAS. Cases and controls were ascertained from a case-control study at the University of Texas MD Anderson Cancer Center conducted between 1997 and 2007 (ref.

3). Cases were newly diagnosed patients with histologically confirmed lung cancer presenting at MD Anderson Cancer who had not previously received treatment other than surgery. Clinical and pathological data were abstracted from patient medical records, and lung cancer histology was coded according to the major histological groups. Specifically, as per ICD-O-2, these groups were SQ: 8070/3; AD: 8140/3, 8250/3, 8260/3, 8310/3, 8480/3, 8251/3, 8490/3. Only patients with predominantly or wholly AD or SQ cancers were included; those with mixed histology or unspecified lung cancers were excluded from the study. Controls were healthy individuals seen for routine care at Kelsey-Seybold clinics in the Houston metropolitan area. Controls were frequency matched to cases according to smoking behavior, age in 5-year categories, ethnicity and sex. Former smoking controls were further frequency matched to former smoking cases according to the number of years since smoking cessation (in 5-year categories). After applying quality controls, data were available on 1,150 cases and 1,134 controls.

Quality control of GWAS data sets.

Standard quality control was performed on all scans, excluding individuals with low call rate (<90%) and extremely high or low heterozygosity ($P < 1.0 \times 10{-}4$), as well as all individuals evaluated to be of non-European ancestry (using the HapMap version 2 CEU, JPT/CHB and YRI populations as a reference; Supplementary Table 1). For apparent first-degree relative pairs, we removed the control from a case-control pair; otherwise, we excluded the individual with the lower call rate.

Replication series.

To validate promising associations from the meta-analysis, we made use of in silico data and imputed genotypes from Harvard and deCODE GWAS data sets together with data from the direct-genotyping Heidelberg-EPIC, ICR, IARC and Toronto replication series.

Harvard. For the Harvard Lung Cancer Susceptibility Study, details of participant recruitment have been described previously50. Replication was based on data derived from 1,000 cases and 1,000 controls genotyped using Illumina HumanHap610-Quad arrays. Cases were patients aged >18 years with newly diagnosed, histologically confirmed primary NSCLC. Controls were healthy non–blood related family members and friends of patients with cancer or with cardiothoracic conditions undergoing surgery. The histological classification of lung tumors was performed by two staff pulmonary pathologists at Massachusetts General Hospital according to ICD-O-3: specifically, AD: 8140/3, 8250/3, 8260/3, 8310/3, 8480/3 8560/3; LCC: 8012/3, 8031/3; SQ: 8070/3, 8071/3, 8072/3, 8074/3; other NSCLC: 8010/3, 8020/3, 8021/3, 8032/3, 8230/3. Unqualified samples were excluded if they fit the following quality-control criteria: (i) overall genotype completion rates <95%; (ii) gender discrepancies; (iii) unexpected duplicates or probable relatives (based on a pairwise identity-by-state value of PI_HAT in PLINK >0.185); (iv) heterozygosity rates >6 times the s.d. from the mean; or (v) individuals evaluated to be of non-European ancestry (using HapMap release 23 including the JPT, CEPH, CEU and YRI populations as a reference). Unqualified SNPs were excluded when they fit the following quality-control criteria: (i) SNPs were not mapped on autosomes; (ii) SNPs had a call rate <95% in all GWAS samples; (iii) SNPs had MAF <0.01; or (iv) the genotype distributions of SNPs deviated from those expected by Hardy-Weinberg equilibrium ($P < 1.0 \times 10{-}6$). After applying these prespecified quality controls, genotype data were available for 984 cases and 970 controls.

deCODE. The Icelandic lung cancer study has been described previously4. The primary source of information on the Icelandic lung cancer cases is the Icelandic Cancer Registry (ICaR), which covers the entire population of Iceland (http://www.cancerregistry.is/krabbameinsskra/indexen.jsp?id=summary). The sources

of data in the ICaR are all pathology and hematology laboratories and all hospital departments and health care facilities in the country. ICaR registration is based on the ICD system and includes information on histology (systemized nomenclature of medicine, SNOMED). ICaR registration also uses the ICD-O system, which takes histology diagnosis into account. Over 94% of diagnoses in the ICaR have histological confirmation. Briefly, according to the ICaR, a total of 4,252 patients were diagnosed with lung cancer from January 1, 1955 to December 31, 2010. Recruitment of both prevalent and incident cases was initiated in 1998, the recruitment is ongoing and DNA samples from lung cancer cases are subjected to whole-genome genotyping as they are collected. The controls used in this study consisted of individuals from other GWAS that were age and sex matched to cases, with no individual disease group accounting for >10% of all controls. Samples were assayed with the Illumina HumanHap300, HumanCNV370, HumanHap610, HumanHap1M, HumanHap660, Omni-1, Omni 2.5 or Omni Express bead chips at deCODE genetics. SNPs were excluded if they had (i) a yield <95%, (ii) MAF < 1% in the population, (iii) deviation from Hardy-Weinberg equilibrium (HWE; P < 10−6), (iv) inheritance error rate (>0.001) or (v) if there was a substantial difference in allele frequency between chip types (in which case the SNP was removed from a single chip type if that resolved the difference, but if it did not then the SNP was removed from all chip types). All samples with a call rate of <97% were removed from the analysis. The Icelandic sample set is drawn from the Icelandic population, which is a small homogeneous founder population with almost no detectable population substructure. Thus, there was no need to adjust for such substructure in the association analysis. In addition, the comprehensive Icelandic genealogy database allowed us to exclude individuals not of Icelandic origin from the analysis. SNP genotypes were phased using the method of long-range phasing51; for the HumanHap series of chips, 304,937 SNPs

were used for long-range phasing, whereas for the Omni series of chips, 564,196 SNPs were used. An initial imputation step was carried out on each chip series separately to create a single harmonized, long-range phased genotype data set consisting of 707,525 SNPs for 95,085 Icelandic individuals. Two sets of genotypes were imputed into this data set with methods previously described52: (i) genotypes for about 38 million variants using the 1000 Genomes phase I integrated variant set (v3) as training set and (ii) genotypes for about 34 million variants identified in 2,230 whole genome–sequenced Icelanders. The first set of imputed genotypes was used for replicating the association with variants in the 5p15.33, 9p21 and 12q13.33 regions using IMPUTE (v2.1.1)53 to perform the case-control analysis. The second set was used when testing the relationship between the p.Lys3326X and c.999del5 genotypes and risk of different cancer types in the Icelandic population using a method that allowed including individuals that had not been chip typed but for whom genotype probabilities were imputed using methods of familial imputation51.

Heidelberg-EPIC. This study comprised 1,253 Heidelberg-EPIC controls and 1,362 lung cancer cases from the Heidelberg lung cancer study recruited between 1994 and 1998 and between 1996 and 2007, respectively. Details of the Heidelberg-EPIC controls and the Heidelberg lung cancer study have been described previously54,55. All subjects were aged 18 years or older, and information on lifestyle risk factors and medical and family history was collected through interviews based on standardized questionnaires. The EPIC Lung and the Heidelberg-EPIC studies were performed independently with no sample overlap with those analyzed as part of the IARC replication series. Histological classification of tumors was obtained from pathology reports, where it was recorded by a staff pulmonary pathologist according to WHO guidelines. Blood samples from patients with malignant lung disease categorized as follows were included: AD, SCLC, NSCLC,

LCC, carcinoid, mixed lung tumors or mixed without SCLC. The above-described EPIC Lung and Heidelberg-EPIC studies were performed independently with no sample overlap. Genotypes for SNPs showed no significant departure from HWE, with the exception of rs13314271 in cases.

ICR replication. This study comprised 2,448 cases (1,664 male; mean age at diagnosis 71.8 years, s.d. 6.7 years) with pathologically confirmed lung cancer ascertained through GELCAPS48 and 2,989 controls (1,469 male; mean age at sampling 60.6 years, s.d. 12.0 years) collected through the National Study of Colorectal Cancer Genetics56 with no personal history of malignancy. Cases were subclassified into histological subtypes based on ICD coding as described above (in the section detailing the ICR GWAS). Both cases and controls were British residents and had self-reported European ancestry. The genotype distributions of genotypes for each of the SNPs typed in replication showed no significant departure from HWE.

IARC replication. This analysis comprised three studies: (i) EPIC Lung2,57, a nested case-control study performed within the EPIC (European Prospective Investigation into Cancer and Nutrition) prospective cohort totaling 1,119 lung cancer cases and 2,546 controls (matched one or two to cases for age, sex, center and time of recruitment) selected from 8 of the 10 countries participating in EPIC (Sweden, Netherlands, UK, France, Germany, Spain, Italy and Norway); (ii) the Szczecin case-control study32, a consecutive series of 849 incident lung cancer cases ascertained from the outpatient oncology clinic in the regional hospital of Szczecin between 2004 and 2007 (the 1,072 controls were individuals without diagnosed cancer or family history of cancer matched to cases by sex, age and region recruited by general medical practitioners); and (iii) Moscow L2, 1,081 newly diagnosed lung cancer cases and 2,119 controls recruited from three hospitals within the Moscow area of Russia between 2007 and 2011. Information

on lifestyle risk factors and medical and family history was collected from subjects by interview using a standard questionnaire. Cases were subclassified into histological subtypes based on ICD-O3 coding as described above (in the section detailing the IARC GWAS). The distributions of genotypes for each of the SNPs typed in replication showed no departure from HWE in each country or study series.

Toronto. This study was conducted in the greater Toronto area from 2008 to 2013. Lung cancer cases were recruited at the hospitals in the network of the University of Toronto. Controls were selected randomly from individuals registered in the family medicine clinics databases and were frequency matched with cases on age and sex. All subjects were interviewed, and information on lifestyle risk factors, occupational history and medical and family history was collected using a standard questionnaire. Tumors were centrally reviewed by the reference pathologist (a member of the International Association for the Study of Lung Cancer (IASLC) committee) and a second pathologist in the University Health Network. If the reviews conflicted, a consensus was arrived at after discussion. Coding of histology was based on 2001 WHO/IASLC. After applying standardized quality control procedures and restricting the data to participants with self-reported European ancestry, data and samples were available on 1,084 cases and 966 controls. The genotype distributions of genotypes for each of the SNPs typed in replication showed no significant departure from HWE.

Replication genotyping.

Genotyping of rs1519542, rs13314271, rs55731496, rs149423192, rs4592420, rs11571833, rs56084662 and rs17879961 was performed using competitive allele-specific PCR KASPar chemistry (LGC, Hertfordshire, UK; UK replication series), Sequenom (Sequenom, Inc., San Diego, US; Toronto replication and Heidelberg-EPIC replication (rs1519542, rs55731496, rs149423192, rs4592420, rs11571833, rs56084662

and rs17879961)) or TaqMan (Carlsbad, CA; IARC replication series and Heidelberg-EPIC replication (rs13314271)). All primers, probes and conditions used are available on request. Call rates for SNP genotypes were >95% in each of the replication series.

To ensure the quality of genotyping in all assays, at least two negative controls and 1–10% duplicates (showing a concordance of >99%) were genotyped at each center. To exclude technical artifacts in genotyping, at the ICR and IARC we performed cross-platform validation of 96 samples and sequenced a set of 96 randomly selected samples from each case and control series to confirm genotyping accuracy. Assays were found to be performing robustly; concordance was >99%.

Statistical and bioinformatic analyses. Data were imputed for all scans for over 10 million SNPs using data from the 1000 Genomes Project (phase 1 integrated release 3, March 2012) as reference using IMPUTE2 v2.1.1 (ref. 53), MaCH58 v1.0 or minimac (version 2012.10.3)59 software (Supplementary Table 1). Genotypes were aligned to the positive strand in both imputation and genotyping. Imputation was conducted separately for each scan in which each GWAS data set was pruned to a common set of SNPs between cases and controls before imputation. As previously described, we set thresholds for imputation quality to retain both potential common and rare variants for validation13,60. Specifically, poorly imputed SNPs defined by an RSQR < 0.30 with MaCH or an information measure Is < 0.40 with IMPUTE2 were excluded from the analyses. Tests of association between imputed SNPs and lung cancer were performed under a probabilistic dosage model in SNPTEST v2.5 (ref. 61), ProbABEL62, MaCH2dat v.124 (ref. 58) or the glm function in R. Principle components generated using common SNPs were included in the analysis to limit the effects of cryptic population stratification that might cause inflation of test statistics. The association between each SNP and lung cancer risk was assessed by Cochran-Armitage trend test. The adequacy of the case-control matching

and possibility of differential genotyping of cases and controls were formally evaluated using Q-Q plots of test statistics. Meta-analysis was undertaken using inverse-variance approaches. The inflation factor λ was based on the 90% least-significant directly typed SNPs63. ORs and associated 95% CIs were calculated by unconditional logistic regression using R (v2.6), Stata v.10 (State College, Texas, US) and PLINK64 (v1.06) software. Cochran's Q statistic to test for heterogeneity and the I2 statistic to quantify the proportion of the total variation due to heterogeneity were calculated65. I2 values ≥75% are considered to be characteristic of large heterogeneity65. Additionally, analyses stratified by histology, sex, age and smoking status (current, former or never) were performed. All statistical tests were two sided.

The fidelity of imputation as assessed by the concordance between imputed and directly genotyped SNPs was examined in a subset of samples from the UK GWAS, MDACC GWAS, IARC GWAS and NCI GWAS discovery series (Supplementary Table 2).

LD metrics were calculated in PLINK using 1000 Genomes data and plotted using SNAP66. LD blocks were defined on the basis of HapMap recombination rate (cM/Mb) as defined using the Oxford recombination hotspots and on the basis of the distribution of CIs defined by Gabriel et al.67.

Relationship between genotypes and smoking.

To examine the relationship between rs11571833 (BRCA2 p.Lys3326X), rs17879961 (CHEK2 p.Ile157Thr) and rs13314271 (TP63) genotype and cigarette consumption (cigarettes per day)68, we used data on 43,693 Icelandic subjects (including 34,850 chip-typed individuals).

Sequence analysis of BRCA2 in constitutional DNA.

At the ICR, targeted sequencing for the BRCA2 mutations c.6275delTT and c.4889C>G was performed by Sanger implemented on an ABI3700 analyzer (Applied Biosystems; primer sequences and conditions are available on request). Mutational analysis of the complete coding region of BRCA2 was based on exome sequencing data generated using Illumina TruSeq capture technology (Illumina, Inc, San Diego, USA). Analysis of Illumina HiSeq2000 (Illumina, Inc, San Diego, USA) sequence data was performed using an in-house pipeline based on the GATK tool kit.

At IARC, Qiagen Generead (SABiosciences/Qiagen Hilde, Germany) was used to amplify the coding region of BRCA2 in rs11571833 heterozygotes.

After library preparation (New England BioLabs, Ipswich, MA, USA), sequencing was performed using an IonTorrent PGM desktop sequencer (Life Technologies, Guilford, San Francisco, CA). Genotypes were called using Ionsuite software. Sequence changes were referenced to the Leiden Open Variation Database (LOVD2) and the BReast CAncer IARC database.

Analysis of TCGA data.

The exomes of 243 individuals with lung SQ and 338 individuals with lung AD in TCGA (Project Number #3230) were analyzed at IARC using an in-house pipeline based on the GATK tool set. Variant calls were annotated using ANNOVAR, making use of use the National Heart, Lung, and Blood Institute's Exome Sequencing Project and 1000 Genomes data.

Copy number variation. Copy number variation was assessed from Human SNP Array 6.0 data. We retrieved level 3 TCGA data comprising normalized log2 ratios of the fluorescence intensities between the target sample and a reference sample. We included only tumor-normal paired data in our analysis. We considered a log2 ratio ≤ 0.5 as

reflecting loss and a log2 ratio >0.5 as reflecting gain. Annotation was performed by adding the genes contained in each of the remaining segments using EnsEMBL databases.

References

38. Omenn, G.S. et al. The β-carotene and retinol efficacy trial (CARET) for chemoprevention of lung cancer in high risk populations: smokers and asbestos-exposed workers. Cancer Res. 54 (suppl. 7), 2038s–2043s (1994).

39. Scélo, G. et al. Occupational exposure to vinyl chloride, acrylonitrile and styrene and lung cancer risk (Europe). Cancer Causes Control 15, 445–452 (2004).

40. Feyler, A. et al. Point: myeloperoxidase −463G→A polymorphism and lung cancer risk. Cancer Epidemiol. Biomarkers Prev. 11, 1550–1554 (2002).

41. Nelis, M. et al. Genetic structure of Europeans: a view from the North-East.

PLoS ONE 4, e5472 (2009).

42. Välk, K. et al. Gene expression profiles of non-small cell lung cancer: survival prediction and new biomarkers. Oncology 79, 283–292 (2010).

43. Holmen, J. et al. The Nord-Trondelag Health Study 1995–97 (HUNT2): objectives, contents, methods and participation. Norsk Epidemiologi 13, 1932 (2003).

44. Landi, M.T. et al. Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer. BMC Public Health 8, 203 (2008).

45. ATBC Cancer Prevention Study Group. The α-tocopherol, β-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. Ann. Epidemiol. 4, 1–10 (1994).

46. Hayes, R.B. et al. Methods for etiologic and early marker investigations in the PLCO trial. Mutat. Res. 592, 147–154 (2005).

47. Calle, E.E. et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. Cancer 94,

2490–2501 (2002).

48. Eisen, T., Matakidou, A. & Houlston, R. Identification of low penetrance alleles for lung cancer: the GEnetic Lung CAncer Predisposition Study (GELCAPS). BMC Cancer 8, 244 (2008).

49. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678 (2007).

50. Su, L. et al. Genotypes and haplotypes of matrix metalloproteinase 1, 3 and 12 genes and the risk of lung cancer. Carcinogenesis 27, 1024–1029 (2006).

51. Kong, A. et al. Parental origin of sequence variants associated with complex diseases. Nature 462, 868–874 (2009).

52. Styrkarsdottir, U. et al. Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. Nature 497, 517–520 (2013).

53. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529 (2009).

54. Boeing, H., Wahrendorf, J. & Becker, N. EPIC-Germany—a source for studies into diet and risk of chronic diseases. European Investigation into Cancer and Nutrition. Ann. Nutr. Metab. 43, 195–204 (1999).

55. Dally, H. et al. The CYP3A4*1B allele increases risk for small cell lung cancer: effect of gender and smoking dose. Pharmacogenetics 13, 607–618 (2003).

56. Penegar, S. et al. National study of colorectal cancer genetics. Br. J. Cancer 97, 1305–1309 (2007).

57. Timofeeva, M.N. et al. Genetic polymorphisms in 15q25 and 19q13 loci, cotinine levels, and risk of lung cancer in EPIC. Cancer Epidemiol. Biomarkers Prev. 20, 2250–2261 (2011).

58. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34, 816–834 (2010).

59. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44, 955–959 (2012).

60. Zeggini, E. et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat. Genet. 40, 638–645 (2008).

61. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11, 499–511 (2010).

62. Aulchenko, Y.S., Struchalin, M.V. & van Duijn, C.M. ProbABEL package for genome-wide association analysis of imputed data. BMC Bioinformatics 11, 134 (2010).

63. Clayton, D.G. et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat. Genet. 37, 1243–1246 (2005).

64. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).

65. Higgins, J.P., Thompson, S.G., Deeks, J.J. & Altman, D.G. Measuring inconsistency in meta-analyses. Br. Med. J. 327, 557–560 (2003).

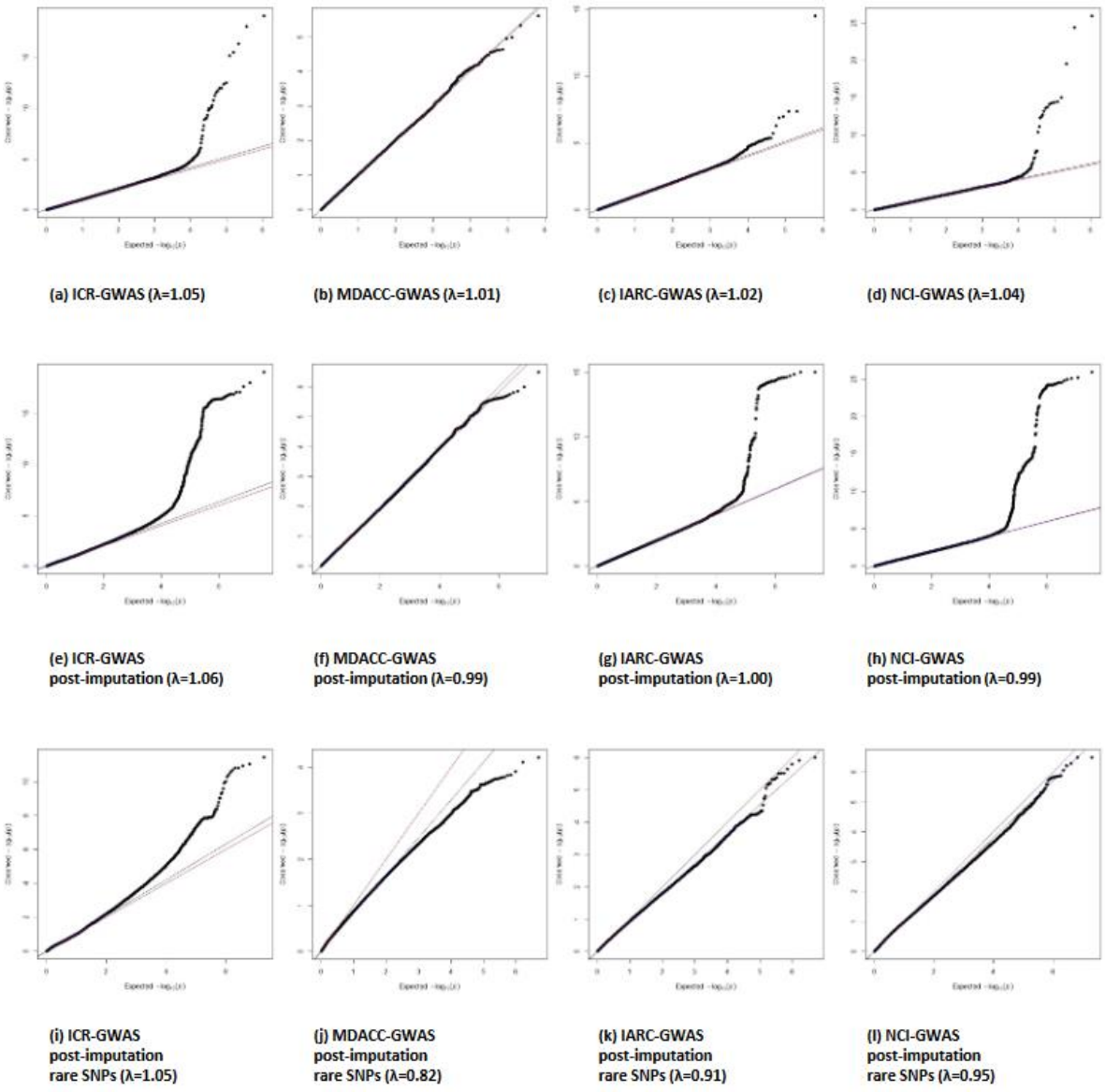
66. Johnson, A.D. et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24, 2938–2939 (2008).

67. Gabriel, S.B. et al. The structure of haplotype blocks in the human genome. Science 296, 2225–2229 (2002).

68. Thorgeirsson, T.E. et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat. Genet. 42, 448–453 (2010).

Supplementary Figure

Supplementary Figure 1 Q-Q plots of Cochran-Armitage trend test statistics for association based on 11,348 cases and 15,861 controls from discovery phase GWASs pre-imputation (a-d); all SNPs post-imputation (e-h) and rare SNPs post-imputation (i-l).



(a) ICR-GWAS (λ=1.05) (b) MDACC-GWAS (λ=1.01) (c) IARC-GWAS (λ=1.02) (d) NCI-GWAS (λ=1.04)

(e) ICR-GWAS post-imputation (λ=1.06) (f) MDACC-GWAS post-imputation (λ=0.99) (g) IARC-GWAS post-imputation (λ=1.00) (h) NCI-GWAS post-imputation (λ=0.99)

(i) ICR-GWAS post-imputation rare SNPs (λ=1.05) (j) MDACC-GWAS post-imputation rare SNPs (λ=0.82) (k) IARC-GWAS post-imputation rare SNPs (λ=0.91) (l) NCI-GWAS post-imputation rare SNPs (λ=0.95)